Point/Counterpoint

Counterpoint: Bias from Population Stratification Is Not a Major Threat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer

Sholom Wacholder,[1] Nathaniel Rothman, and Neil Caporaso

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20854

Introduction

In their Point, Thomas and Witte (1) forcefully lay out their case for seriously considering the impact of population stratification in so-called "association" studies of genetic factors and cancer. In our Counterpoint, we discuss the nature of population stratification; the conditions that are necessary for it to occur and when they are likely in studies of cancer; the evidence about whether important bias or excess false-positive findings are consequences of population stratification; how we might determine empirically the seriousness of the consequences; and possible solutions to the problem, if it exists, including replication, use of genomic control, and use of related controls. We conclude with a statement of our view of the current situation and where it differs from that of Thomas and Witte (1). Specifically, we affirm our view, enunciated in work published previously (2), that population stratification is not a serious threat to the reliability of cohort and case-control studies of cancer, at least in studies of non-Hispanics of European descent, with unrelated controls; the main reason for the restriction to studies of Europeans is the lack of solid empirical data in other groups.

Population Stratification: What Is It and When Will It Cause Important Bias?

Population stratification refers to a particular form of confounding. In cohort and case-control studies of genetic variants, the bias from population stratification is the distortion in the value of an observed association between the genetic variant G and disease D that can occur when G is associated with some true risk factor E that varies by ethnicity (Fig. 1). In population stratification, risks of disease in related individuals, particularly those in the same ethnic group, are more similar than risks of disease in more distantly related or completely unrelated individuals. The homogeneity may be attributable to similar lifestyle or to greater similarity in the presence of one or more risk-conferring alleles. Ethnicity *per se* does not explain the risk; it is only a marker for individuals at similar risk. Similarly, a gradient in risk of disease by socioeconomic status does not itself cause disease but may be a reflection of similarity in lifestyle or access to preventive health care. Whether the consideration is ethnicity or socioeconomic status, controlling for the factors that explain the similarity in risk eliminates any bias.

What is the impact of population stratification? We acknowledge, without reservation, that population stratification exists and causes bias. We note, however, that the term "bias," refers both to the existence of some distortion of the measure of effect and to the amount of distortion. In an epidemiological study, it is not the existence of bias that is intrinsically troublesome but its potential magnitude (3); otherwise, we would not do studies with, for example, incomplete case ascertainment, moderate response rates in controls, or exposures measured with error. Our criterion for evaluating the consequences of population stratification is quantitative: it is the potential to have enough bias in a study so that its conclusions and interpretation change materially.

Several conditions must be met before there will be substantial bias in a cohort or case-control study designed to quantify the effects of common polymorphisms on the risk of cancer at a given site (2):

(*a*) There must be substantial variation across ethnicities in the frequency of the variant genotype being considered.

(*b*) There must be substantial variation across ethnicities in disease rates after adjustment for risk factors, other than the genotype of interest, that were collected in the study; typically, but certainly not always, adjustment will reduce the interethnic differences in cancer rates.

(*c*) The allele frequencies must track with the adjusted cancer rates across ethnicities, for reasons other than the effects of the allele of interest (2). For example, an allele with a clade or gradient of increasing frequency from North to South in Europe might track with a factor that affects cancer risk, such as consumption of beer and butter rather than wine and olive oil (2); this could induce bias from population stratification when studying the effect of the allele.

(*d*) Collection of ethnic information from study participants must not be able to reduce bias to an acceptable level.

Population stratification does not occur in an ethnically homogeneous population. Nonetheless, we showed that the potential for bias is greatest with two or three ethnicities and tends to diminish as the number of ethnicities increases; the bias tends to be less severe in a more diverse study population (2). Even if there is substantial bias in a single study, it is highly unlikely that bias in the same direction will occur in a second study in a population with a different ethnic mix (2).

Received 3/12/01; revised 9/20/01; accepted 10/4/01.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] To whom requests for reprints should be addressed, at Division of Cancer Epidemiology and Genetics, National Cancer Institute, EPS 8046, Bethesda MD 20892-7244 (for courier delivery: 6120 Executive Blvd., Rockville MD 20854). Phone: (301) 496-3358; Fax: (301) 402-0081; E-mail: Wacholder@nih.gov.

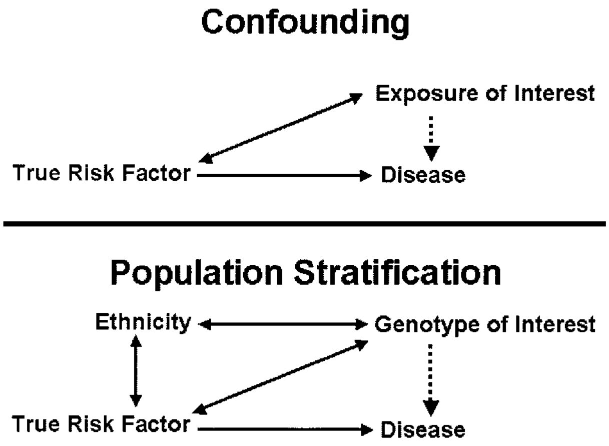

Fig. 1. Classical confounding and population stratification. In population stratification, the frequency of an unmeasured risk factor for disease differs by ethnicity. Broken lines with arrow indicate an association that is potentially confounded by the true risk factor. Solid unidirectional arrows indicate the direction of causal relationship. Solid bidirectional arrows indicate a correlation that may or may not be causal. Reprinted from *Journal of the National Cancer Institute (1).*

Under What Scenarios Is Important Bias from Population Stratification Likely?

Calculations. Theoretical results in Table 1, calculated exactly like those accepted by Thomas and Witte (1), show that a bias factor for the rate ratio, and by implication, the odds ratio and the risk ratio, >15% attributable to population stratification is unlikely when multiple ethnic groups are considered, unless the range of cancer rates in the ethnic groups is at least 2-fold and the frequency of the at-risk genotype varies substantially. These bias factors, which do not depend on the strength of the relationship between the allele being studied and disease, should be considered acceptable for epidemiological studies, especially when considered against realistic alternative study designs. But even the small levels of bias in our work are exaggerated:

(*a*) They are calculated under the assumption that there is no useful information available on either the true risk factor driving the differences in rates or on ethnicity itself and that classical methods for the reduction of bias from confounding are not used.

(*b*) The extreme values shown are obtained when allele frequency and cancer rates for those without the at-risk genotype are strongly correlated.

(*c*) Furthermore, it seems safe to assume that the variation in cancer rates in ethnic groups in the United States is less than the variation in rates in their ancestors' countries of origin for two reasons: (*i*) cancer rates in descendants of immigrants become closer to those of long-term residents of their new homeland (4); and (*ii*) interethnic variation in allele frequencies and cancer rates are dampened by intermarriage among these groups.

Effect of Differences in Distribution of Risk Factors on Magnitude of Ratios in Cancer Rates. To orient ourselves in Table 1, we need to consider the magnitude of the difference in rate of cancer in different groups induced by the difference in the distribution of one or more risk factors in the groups. Thomas and Witte (1) mention the "enormous risk attributable to genes" in their conclusion. Their concern seems to be that if a large fraction of cases are attributable to variation at a single locus with high penetrance, an allele not causally related to risk may be associated with disease because of population stratification; the geographic variation in rates would then closely follow the distribution of the at-risk allele or genotype, and a study allele that is associated (whether because of founder effects, linkage disequilibrium, or any other reason) with the genotype conferring risk would appear associated with cancer. In other words, this is an example of classical confounding.

However, it is the magnitude of the attributable risk, not the relative risk or the penetrance, of a risk factor that directly determines the variation in disease rates across groups, and hence the confounding potential of the risk factor. Attributable risk can be defined as $AR = (I_C - I_0)/I_0$, where I_C is the crude incidence rate in the population (calculated without taking level of exposure into account), and I_0 is the rate in the unexposed; after rearranging terms, one sees that $I_C/I_0 = 1/(1 - AR)$. Thus, in a population with an attributable risk from mutations in a specific gene of 33%, the crude incidence rate is only 1.5-fold higher than the risk in a population that is completely unexposed. The highest attributable risk from alleles for a relatively common adult cancer that we know is due to the effects of *BRCA1* and *BRCA2* founder mutations in Ashkenazi Jewish women. Despite the very large relative risk of >20 and an extremely high carrier frequency of 1.7% for high-penetrance founder mutations (5), the overall rate of ovarian cancer in Jewish women would be only 1.4 times the rate in an ethnic group with no mutations; Table 1 shows that this is far too small of an effect to cause much bias. In general, a highly penetrant allele is not sufficient to cause major bias; the allele must be frequent for the attributable risk to be high. But a very high attributable risk from a gene is not to be expected often for complex diseases such as adult cancer.

Moreover, when a disease is caused independently by alleles in several different genes not close together on the same chromosome, the possibility of important bias from population stratification is greatly reduced, even if each of the individual alleles confers high penetrance. No candidate allele could be strongly associated with most of these true-risk alleles unless the risk alleles were in linkage disequilibrium with one another; instead, typically, there would be a mixture of positive and negative correlations leading to considerably less bias than if the variation in risk were attributable to differing frequency of a single allele.

Importance of Environmental Factors. The descriptive epidemiology of adult cancer, particularly geographic and temporal variation and studies of migrants, and analytic studies lead us to infer (6, 7) that international differences in cancer rates are determined more by environmental, cultural, and behavioral factors, perhaps modified by genes, than by genes acting alone. If our view is correct, then the international and ethnic variation in cancer rates are unlikely to be explained solely by alleles at one or more loci in linkage disequilibrium with a candidate allele simply because of founder effects. Furthermore, with the possible exception of genes that may influence a behavioral exposure, such as *ALDH* alleles and alcohol (8), tracking of cancer rates and frequency of high-risk genotypes across ethnicities, a necessary condition for bias, is likely to be minimal; even if tracking is present in one population, similar tracking is highly unlikely to be found in a replication of the study in a population consisting of different ethnicities.

Some might argue with our premise (9), although Thomas and Witte (1) do not. The totality of the evidence from family and twin studies appears to be consistent with various numbers of causal genes with a wide range of attributable risk (10). However, the well-documented environmental determinants

Table 1 Bias factors calculated under different hypothetical variation in cancer incidence rates and high-risk genotype frequency[a]

Risk ratio[b]	Frequency of at-risk genotype[c]		Bias factor among all 28 subsets of size 2 among the eight groups		Bias factor among all 40,320 orderings of the eight groups											
Max	Min	Max	Min	Max	Min	2.5 %ile	5 %ile	10 %ile	25 %ile	50 %ile	75 %ile	90 %ile	95 %ile	97.5 %ile	Max bias	Mean bias
1.0	0.1	0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.0	0.2	0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.0	0.3	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.1	0.1	0.3	0.97	1.03	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.00
1.1	0.2	0.5	0.97	1.03	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.00
1.1	0.3	0.9	0.94	1.06	0.97	0.98	0.98	0.99	0.99	1.00	1.01	1.01	1.02	1.02	1.03	1.00
1.2	0.1	0.3	0.94	1.06	0.98	0.98	0.98	0.99	0.99	1.00	1.01	1.01	1.02	1.02	1.02	1.00
1.2	0.2	0.5	0.94	1.06	0.97	0.98	0.98	0.99	0.99	1.00	1.01	1.01	1.02	1.02	1.03	1.00
1.2	0.3	0.9	0.89	1.12	0.95	0.97	0.97	0.98	0.99	1.00	1.01	1.02	1.03	1.04	1.05	1.00
1.3	0.1	0.3	0.92	1.08	0.97	0.98	0.98	0.98	0.99	1.00	1.01	1.02	1.02	1.03	1.04	1.00
1.3	0.2	0.5	0.92	1.09	0.96	0.97	0.98	0.98	0.99	1.00	1.01	1.02	1.02	1.03	1.04	1.00
1.3	0.3	0.9	0.85	1.18	0.93	0.95	0.96	0.97	0.98	1.00	1.02	1.04	1.04	1.05	1.07	1.00
1.4	0.1	0.3	0.90	1.11	0.96	0.97	0.97	0.98	0.99	1.00	1.01	1.02	1.03	1.03	1.05	1.00
1.4	0.2	0.5	0.89	1.11	0.95	0.97	0.97	0.98	0.99	1.00	1.01	1.02	1.03	1.03	1.05	1.00
1.4	0.3	0.9	0.81	1.24	0.92	0.94	0.95	0.96	0.97	1.00	1.03	1.05	1.06	1.07	1.09	1.00
1.5	0.1	0.3	0.88	1.13	0.95	0.96	0.97	0.97	0.98	1.00	1.02	1.03	1.03	1.04	1.05	1.00
1.5	0.2	0.5	0.87	1.14	0.94	0.96	0.97	0.97	0.98	1.00	1.02	1.03	1.04	1.04	1.06	1.00
1.5	0.3	0.9	0.78	1.29	0.90	0.93	0.94	0.95	0.97	1.00	1.03	1.06	1.07	1.08	1.11	1.00
1.6	0.1	0.3	0.86	1.15	0.94	0.96	0.96	0.97	0.98	1.00	1.02	1.03	1.04	1.04	1.06	1.00
1.6	0.2	0.5	0.86	1.16	0.94	0.95	0.96	0.97	0.98	1.00	1.02	1.03	1.04	1.05	1.07	1.00
1.6	0.3	0.9	0.75	1.35	0.88	0.92	0.93	0.94	0.97	1.00	1.04	1.06	1.08	1.09	1.13	1.00
1.7	0.1	0.3	0.84	1.17	0.93	0.95	0.96	0.97	0.98	1.00	1.02	1.03	1.04	1.05	1.07	1.00
1.7	0.2	0.5	0.84	1.18	0.93	0.95	0.96	0.96	0.98	1.00	1.02	1.04	1.05	1.05	1.08	1.00
1.7	0.3	0.9	0.73	1.40	0.87	0.91	0.92	0.93	0.96	1.00	1.04	1.07	1.09	1.11	1.15	1.00
1.8	0.1	0.3	0.83	1.19	0.92	0.95	0.95	0.96	0.98	1.00	1.02	1.04	1.05	1.06	1.08	1.00
1.8	0.2	0.5	0.82	1.20	0.92	0.94	0.95	0.96	0.98	1.00	1.02	1.04	1.05	1.06	1.08	1.00
1.8	0.3	0.9	0.71	1.45	0.86	0.90	0.91	0.93	0.96	1.00	1.04	1.08	1.10	1.12	1.17	1.00
1.9	0.1	0.3	0.81	1.20	0.92	0.94	0.95	0.96	0.98	1.00	1.02	1.04	1.05	1.06	1.08	1.00
1.9	0.2	0.5	0.81	1.22	0.91	0.94	0.95	0.96	0.98	1.00	1.03	1.04	1.06	1.06	1.09	1.00
1.9	0.3	0.9	0.69	1.51	0.85	0.89	0.90	0.92	0.95	1.00	1.05	1.09	1.11	1.13	1.18	1.00
2.0	0.1	0.3	0.80	1.22	0.91	0.94	0.95	0.96	0.97	1.00	1.03	1.05	1.06	1.06	1.09	1.00
2.0	0.2	0.5	0.80	1.24	0.91	0.93	0.94	0.95	0.97	1.00	1.03	1.05	1.06	1.07	1.10	1.00
2.0	0.3	0.9	0.67	1.56	0.84	0.88	0.90	0.92	0.95	1.00	1.05	1.09	1.12	1.14	1.20	1.00
2.1	0.1	0.3	0.79	1.23	0.91	0.93	0.94	0.95	0.97	1.00	1.03	1.05	1.06	1.07	1.10	1.00
2.1	0.2	0.5	0.78	1.25	0.90	0.93	0.94	0.95	0.97	1.00	1.03	1.05	1.06	1.07	1.10	1.00
2.1	0.3	0.9	0.65	1.60	0.83	0.87	0.89	0.91	0.95	1.00	1.06	1.10	1.13	1.15	1.21	1.00
2.2	0.1	0.3	0.78	1.25	0.90	0.93	0.94	0.95	0.97	1.00	1.03	1.05	1.06	1.07	1.10	1.00
2.2	0.2	0.5	0.77	1.27	0.90	0.93	0.94	0.95	0.97	1.00	1.03	1.05	1.07	1.08	1.11	1.00
2.2	0.3	0.9	0.63	1.65	0.82	0.87	0.88	0.91	0.94	1.00	1.06	1.11	1.13	1.16	1.23	1.00
2.3	0.1	0.3	0.77	1.26	0.90	0.93	0.94	0.95	0.97	1.00	1.03	1.05	1.07	1.08	1.11	1.00
2.3	0.2	0.5	0.76	1.29	0.89	0.92	0.93	0.95	0.97	1.00	1.03	1.06	1.07	1.08	1.12	1.00
2.3	0.3	0.9	0.62	1.70	0.81	0.86	0.88	0.90	0.94	1.00	1.06	1.11	1.14	1.17	1.24	1.00
2.4	0.1	0.3	0.76	1.27	0.89	0.92	0.93	0.95	0.97	1.00	1.03	1.06	1.07	1.08	1.11	1.00
2.4	0.2	0.5	0.75	1.30	0.89	0.92	0.93	0.94	0.97	1.00	1.03	1.06	1.07	1.09	1.12	1.00
2.4	0.3	0.9	0.61	1.74	0.81	0.86	0.87	0.90	0.94	1.00	1.07	1.12	1.15	1.17	1.25	1.00
2.5	0.1	0.3	0.75	1.28	0.89	0.92	0.93	0.94	0.97	1.00	1.03	1.06	1.07	1.08	1.12	1.00
2.5	0.2	0.5	0.74	1.31	0.88	0.92	0.93	0.94	0.97	1.00	1.04	1.06	1.08	1.09	1.13	1.00
2.5	0.3	0.9	0.59	1.79	0.80	0.85	0.87	0.89	0.94	1.00	1.07	1.12	1.16	1.18	1.27	1.00

[a] The table entries are confounding risk ratios, calculated according to the formula provided by Wacholder *et al.* (2). They show the bias factor when there are two or eight equally common ethnic groups in the population, and no attempt has been made to adjust for ethnicity or other risk factors.

[b] Ratio of the highest to the lowest risk of disease in the eight ethnic groups. The eight rates are spaced equidistantly on a linear scale.

[c] Minimum and maximum frequency of at-risk genotype in the eight ethnic groups. The eight frequencies are spaced equidistantly on a linear scale.

and the rarity of striking familial aggregation of many cancers argue against a single common, high-penetrant allele being responsible for a large fraction of cancer, except for the rarest forms such as retinoblastoma.

Variability in Allele Frequency Is Not Sufficient to Cause Population Stratification. Sometimes another of the conditions for important bias from population stratification, substantial variation in frequency of the study allele across ethnicities, is met; for example, as Thomas and Witte (1) note, *HLA* alleles have considerable interethnic variation in frequency. We agree that special attention should be paid in this situation but also point out that the absence of one or more of the other conditions eliminates or dramatically reduces the bias. For example, Risch (11) has observed that "associations between specific *HLA* antigens and a variety of diseases (mostly autoimmune) have been reported and recently confirmed—for example with insulin-dependent diabetes mellitus, multiple sclerosis, rheumatoid arthritis, psoriasis, celiac disease, narcolepsy, hemochromatosis, and many others." Many of the confirmed associations between *HLA* antigens and these conditions are based on studies with unrelated controls. Furthermore, basic analytic principles can minimize bias (12).

Some other concerns raised by Thomas and Witte (1) are alleviated by considering our theoretical results more closely. The presence of population heterogeneity in allele frequencies and in disease rates are not sufficient to cause bias; as they concede, major bias from population stratification requires tracking of allele frequencies with disease rates and failure to control for ethnicity through tools such as self-report. It follows that except for HLA and other alleles with wide interethnic variation and diseases with large ethnic variation in rates, there is no need for serious concern. Furthermore, when race can be determined easily or when cancer rates do not vary greatly among races, the presence of multiple races (as in California) cannot be a major problem. Allele frequencies and cancer rates from mixed-race individuals are likely to be intermediate between those of the races of their ancestors; these lead to less potential for bias, even if arbitrary racial categorization is used, than in a population that was a mixture of endogamous groups (2). Finally, one or more areas of very high or very low disease rate does not cause major bias unless it represents a large proportion of the total population and it is characterized by an extreme allele frequency.

Cryptic Correlation. Thomas and Witte (1) raise the related issue of so-called cryptic correlation (13), which refers to homogeneity of risk attributable to shared genetic background of distantly related individuals. We know that cryptic correlation can lead to inappropriately narrow confidence intervals, but its quantitative impact depends on the importance of genetic factors on disease. In fact, unmeasured environmental factors can affect significance levels in exactly the same way. If empirical studies show that these do have a major impact on significance tests and confidence intervals, perhaps reduction in the standard 0.05 critical value is called for. Still, the impact is likely to be less than Thomas and Witte (1) imply. Even if the overall type I error rate of 5% were reduced from 1.96 to 2.17 (the critical value for nominal 3% size), the required sample size accounting for 10,000 multiple comparisons via Bonferroni correction would increase by only ~5% when power is 80%.

Is Population Stratification Responsible for Failures to Replicate Findings?

Is there convincing evidence that any specific false-positive finding in the literature is a consequence of population stratification or that population stratification increases the false-positive rate in the aggregate in any substantial way? Thomas and Witte (1) discuss this point at length but finally indicate that they would answer negatively.

We recognize that there is a perception that a large fraction of initial published reports of allele-disease associations have positive findings and that many of these positive findings are eventually refuted. But is the perception grounded in fact and is population stratification responsible? We should consider three specific questions:

(*a*) Is the percentage of published false-positive findings from cohort and case-control studies based on unrelated controls higher than from similar studies using related controls?

(*b*) Do population studies of genetic polymorphisms with unrelated controls have a higher rate of false positives than studies of other factors or even of linkage studies?

(*c*) Is population stratification responsible for many of the "excess" false positives, if there is an excess?

These questions have not been addressed empirically. Given publication bias and self-censorship, the task might be impossible. But if the answers to these intriguing questions are positive, then population stratification indeed is a serious problem.

There are several plausible alternative explanations for false-positive reports in studies with unrelated controls:

(*a*) Many studies, particularly in the early years of studying the effects of genes on cancer, have used poor epidemiological designs and protocols. False-positive reports can be a consequence of using clinic patients as cases and samples of convenience such as medical students or laboratory technicians as controls. Although the failure to consider any differences in race or ethnicity in these studies might lead some to attribute the problem to population stratification, we would consider the studies to violate basic design principles, even if there was matching on ethnicity. We share the view of Morton and Collins (14) that it is facile to attribute the failure to replicate a population study to population stratification rather than poor design; the solution is better adherence to the principles of epidemiological studies for control selection (3), not necessarily using related controls.

(*b*) Undoubtedly, associations between most common alleles and specific diseases are null. The magnitudes of the effects of the truly positive ones investigated in population-based studies are small, as for *NAT2* and bladder cancer (15), and often the studies are small. Consequently, power is low. The positive predictive value of positive reports for true causal relationships increases with power and high prior likelihood of being true (16); a high proportion of false-positive reports is inevitable when testing unlikely hypotheses, especially with low power (17). Furthermore, the power of the replication studies is often low as well; therefore, some failures to replicate can be attributed to false negatives in the confirmatory studies.

(*c*) Studies of complex disease like cancer are intrinsically more difficult than linkage or other studies of a disease caused exclusively by a single fully penetrant gene. Case-control studies of a common polymorphism with low penetrance and low attributable risk for cancer are intrinsically more susceptible to false positives than studies of a monogenic disease caused by a single allele (18). The obstacles inherent in studying complex diseases cannot be avoided because common, low-penetrance polymorphisms may account for a substantial number of cases of disease.

(*d*) Multiple comparisons; data dredging, for example, by subgroup analysis and allowing for dominant and recessive modes of inheritance; and *P*-value creep, *i.e.*, rejecting when the *P*-value is slightly or not-so-slightly above the nominal standard, also increase the apparent frequency of false positives.

(*e*) The failure to replicate a finding from a study of an allele in another population can be a function of different distributions of an unmeasured environmental exposure that interacts with the variant to cause cancer.

(*f*) The very fact that case-control studies are often easy to replicate in existing cohort or other case-control studies increases the speed at which false positives can be identified and repudiated, compared with other designs where organizing a study is time-consuming, such as those requiring multiple families.

Has Population Stratification Been Shown to Create Specific Misleading Results?

Thomas and Witte (1) call several candidate gene associations "classic" examples of population stratification. None of the examples provide a convincing argument against the credibility of results of high-quality epidemiological studies with unrelated controls, and none is a demonstration of population stratification misleading the scientific community. The salutary lesson from the Pima-Indian study (12) is that statistical adjustment can reduce or eliminate a large bias even in an extreme situation, not that population stratification is a ubiquitous concern, as some have concluded. In the other examples, there are "culprits" at least as likely as population stratification for the lack of replication. Differences in findings on *DRD2* could be related to control selection, differences in access or availability of alcohol, presence of other addictive disorders in cases and controls, or differences in phenotype or case definition, including whether it is defined by a strict quantitative threshold on reported consumption of alcohol alone or more broadly and inconsistency across ethnicities in who is labeled an alcoholic. The insulin-dependent diabetes mellitus example is an argument for a more benign view of the effects of population stratification that would make unnecessary the wait for positive findings from insensitive sib-pair studies after positive results from unrelated controls "across several populations." Their final example is a cancer study in African Americans, where we are more agnostic because of lack of data, difficulties in self-report of ethnicity, and admixture; however, an alternative explanation of the discrepant findings might be lack of comparability in socioeconomic status and healthiness of lifestyle between cases who are seen for treatment of prostate cancer and controls in a screening program at the same hospital. We also agree with Thomas and Witte's (1) point that there "remain some questions surrounding [use of] genomic control" to identify instances of important population stratification.

Thomas and Witte (1) cite the reports of Terwilliger and Weiss (19) and of Ioannidis *et al.* (20) to suggest that failure to replicate might be a serious problem. We do not agree that these provide salient evidence about the importance of population stratification. Terwilliger and Weiss (19) find that reported *P*-values are consistent with a global null hypothesis of no relationship between alleles and psychiatric diseases, *i.e.*, that the empirical evidence does not support the existence of any true associations, much less an excess of false-positive findings attributable to population stratification. They themselves blame the failure to replicate on "investigators too frequently gambling on and publishing results in situations when the evidence is not at all compelling." Ioannidis' *et al.* (20) finding might be, at least in part, a regression to the mean phenomenon; if the first finding is noted because a test crosses a threshold, one would expect subsequent findings to be less likely to cross the threshold.

What Can Be Done to Determine Empirically Whether Population Stratification Is Likely to Be a Major Source of Bias?

Thomas and Witte (1) propose additional empirical work, especially in African Americans and other groups of non-European origin, to complement our first efforts (2, 21) and resolve the issue to everyone's satisfaction. It is important to identify the small (or null) subset of settings where population stratification cannot be handled by questionnaire data, standard epidemiological methods, or perhaps genomic control in studies using unrelated controls. In the meantime, however, we think that our theoretical demonstration on the impact of population stratification indicates that the bias is not very substantial, except in extreme cases, which should usually be easy to identify. We are wary, therefore, of the call of Thomas and Witte (1) for a "systematic program of research" if it will impede the funding and publication of perfectly good study proposals, including those based on existing cohorts (22) and studies using unrelated controls. In fact, we support adding to the agenda research on biases from studies with related controls.

Given the failure of empirical studies to provide strong evidence of major bias or large numbers of false positives in studies with unrelated controls, how can we tell whether population stratification is a major problem? We have a specific proposal for additional empirical work using the same unlinked markers that Thomas and Witte (1) and others want to use for genomic control. We would characterize distinct ethnic groups in cohort and population-based case-control studies by combinations of markers that do not cause disease themselves, are not in linkage disequilibrium with alleles that do cause disease, and vary strongly by ethnicity. If population stratification is a serious problem, there will be observable gradients in cancer risk, *i.e.*, variability in rate ratios, as in Table 1, by genomically defined ethnicities, sufficient to affect the bias factor, after adjustment for self-reported ethnicity and known risk factors; if not, population stratification is of less concern.

What Are Possible Solutions If Population Stratification Does Indeed Cause Serious Bias?

Replication in Different Populations. What can one do when and if it turns out that population stratification is a problem? The first solution is replication. Major bias in the same direction in populations with substantially different ethnic mixes is very unlikely because the conditions that allowed major bias are unlikely to be repeated (2).

Genomic Control. Thomas and Witte (1) discuss the possible use of genomic control of population stratification. Genomic control uses markers unrelated to disease to correct the bias (13, 23–25). We believe that these methods can sometimes provide an inexpensive and easy way to enhance the credibility of studies with unrelated controls by ruling out even the remote possibility of bias. Perhaps the genetic approaches will prove particularly useful when ethnic differences are important and self-report is very difficult, as for African Americans, from whom we do not know how to capture information on ethnic ancestry or their degree of European ancestry via questionnaire. By contrast, among Hispanic or Asian immigrants to the United States, it is relatively easy to determine broad geographic origin (Cuban, Mexican, Puerto Rican, Salvadoran; Chinese, Filipino, Korean, Japanese, Vietnamese).

We doubt, however, that adjustment for genomically determined ethnicity will often have much effect on estimates of association, especially in non-Hispanic Europeans. Confounding by interethnic differences in disease risk will not be eliminated by a genetic definition of ethnicity if the main sources of ethnic variation in cancer rates are not genetic. Race or ethnicity is more than just a function of genes, and the racial and ethnic differences in disease rates are more likely explained by behavioral, cultural, and socioeconomic factors, as one can infer from migrant studies (4). Therefore, genomic determination of ethnicity is probably not the most effective way to control for a determinant of risk that is a function of social, economic, behavioral, cultural, environmental, and religious characteristics, as well as access to health care.

We believe, instead, that self-reported ethnicity, which is probably more highly correlated with cultural and behavioral factors than genome-based ethnicity, may be a better and more appropriate tool to reduce confounding in cancer studies in European Americans, and perhaps other groups, where population stratification is a concern. Lin and Kelsey (26) discuss tools for improving collection of ethnic and racial information from questionnaires (but not genetic panels); they point out that no single measure will capture the important sources of variation in disease rates. However, we do not agree with the assertion of Thomas and Witte (1) that fine grouping of ethnic origin is needed, and mixed-ethnicity individuals must be "treated appropriately," given the small bias when no ethnic adjustment is made (Table 1). Furthermore, there could be a loss from overmatching from a possibly higher correlation between genetically determined ethnicity and the allele of interest, leading to loss of power.

If indeed differences in cancer rates are to a large extent manifestations of culturally mediated behavioral differences, one would expect that self-report of ethnicity would, in principle, be accurate, and as Thomas and Witte (1) concede, would be more useful than genomic control. An ethnic group is unlikely to maintain its own cultural practices unless its members are able to distinguish themselves from others. Thus, for example, in Lander and Schork's (18) classic example of association between genes for the chop-stick phenotype, standard epidemiological practice would enable an investigator to distinguish between San Franciscans of Chinese and European ancestry and adjust in the analysis.

Thus, in our view, genetic markers of ethnicity are unlikely to create a better proxy than self-reported ethnicity for cultural practices that are not known or easy to determine but strongly affect cancer risk. We fail to see how the best adjustment for ethnicity based on a genetic panel should do a better job than urban/rural, socioeconomic status, occupation, education, and/or self-reported ethnicity. If, on the other hand, there is some common allele with a substantial relative risk and a high population attributable risk that explains the differences in rates, and there is no epidemiological tool that can capture ethnicity well, a set of markers that discriminates among ethnic groups might reduce the confounding bias. We accept that genetic markers may be helpful in studies of unrelated controls, particularly when the variation in rates is attributable to one or a few genetic factors or reliable self-report of ethnicity is not available. We doubt, however, that they are a panacea, especially if they lead to reduced power and thereby to increased frequency of false positives.

Is Replication with Related Controls Necessary to Establish a Causal Relationship?

Rationale for and Difficulties from Use of Related Controls. Studies with related controls become more valuable if studies with unrelated controls have serious problems, but they are hardly a gold standard. Reliance on related controls has serious practical and efficiency limitations and potentially important biases in studies of cancer. In a disease of old age, such as many cancers, siblings, much less parents, are likely to be unavailable, especially for those that are mostly or completely confined to one sex, particularly men. This problem is likely to be worse for ethnic groups with recent migration to the area where the study is taking place (27). Availability of a sibling may be associated with fertility, birth order, other causes of mortality, socioeconomic and occupational status related to residential mobility, vital status of the case, or social factors that determine sibling relationships; one or more of these might be related to the exposure of interest, with bias as a consequence.

There are also power implications from using related controls. Teng and Risch (28) report that there can be a substantial increase in power for studying genetic effects from using unrelated controls compared with unaffected siblings. For example, when using pooling in a realistic setting (power $= 1 - \beta = 0.8$, size $\alpha = 5 \times 10^{-8}$) of simplex families with a dominant risk model with relative risk of 4 and an allele frequency of 20%, choosing two unrelated controls per case requires 158 cases, whereas the corresponding study with two unaffected sibs for each case requires 357 cases. For duplex families (2 affected siblings), the corresponding advantage is even greater: 73 *versus* 247. Thus, choosing unrelated controls can be much more powerful than choosing siblings for studies of a genetic effect (28). Furthermore, overmatching on environmental variables that aggregate in families can reduce the power for studying environmental exposures. This is a particular problem when assessing the effect of a behavior during childhood or a practice begun before adulthood, such as smoking.

In addition, there are some unresolved methodologic and fieldwork issues that may reduce the usefulness of using sibling controls in studies designed to study the effects of both genetic and environmental factors. There can be bias in studying environmental variables with levels that vary geographically. If all of the cases in a study are identified from an area with high levels of environmental pollution, then siblings or other relatives who do not live in the area will tend to have less exposure, leading to falsely elevated estimates of the main effects and distorted estimates of interaction. This bias reflects the violation of the control selection principle that requires controls to be considered as cases if they would develop disease (3); the low-exposure siblings who live elsewhere can be controls but have no chance to become a case. Requiring all controls to be from the study area solves the bias problem but at a high cost; cases without siblings in the area would also need to be excluded.

Some studies, of course, may be improved by including sibling controls, if they are complementary to or more appropriate, overall, than unrelated controls; however, the extra financial cost, logistical challenges, and potential bias from using study designs with parental, sib, and cousin controls in populations need to be considered for cancer in both young and old people. Related controls are totally impractical for some studies of important environmental factors with clinical as well as etiologic relevance, such as genetic modification of the toxicity and efficacy of chemotherapeutic drugs. On the other hand, related controls are useful for studying the interaction between an environmental factor and a rare allele, as noted by Witte *et al.* (27). For example, the Israeli ovarian cancer study (5) mentioned above examined the effects of parity and use of oral contraceptives in women who were *BRCA1* and *BRCA2* carriers. In reflection of the high relative risk for ovarian cancer from these mutations, 29% of cases but only 1.5% of controls carried one of the three studied founder mutations. With sibling controls, there would be many more carrier controls with which to address the question directly. However, the rationale for related controls here is power, not protection against population stratification.

Summary

In earlier work, we argued that bias from population stratification is unlikely to be substantial in studies of cancer in non-Hispanic Europeans (2). Thomas and Witte (1) agree, stating "We have no fundamental quarrel with these conclusions." Yet they suggest that "no candidate gene association should be considered 'established' until confirmed by a [family-based case-control study], or at least by multiple well-designed studies in different populations where any effects of population stratification or other methodologic biases are unlikely to act in a consistent manner," apparently even where they agree that bias from population stratification is likely to be minimal. We certainly support the need to replicate epidemiological findings; we noted previously that in a worst-case scenario where all of the necessary conditions for population stratification are met in a single study, "the results will probably not be replicated because the same conditions are unlikely to exist [in the second study]" (2). We reject the principle of the use of related controls as a universal gold standard. They are far from perfect themselves, and provide costly insurance only against the bias from population stratification, which is likely to be small in most instances.

We find no justification for this argument other than an appeal to the authority of a *Nature Genetics* editorial (29) from which they quote at length. The editorial does state that submissions should have associations "observed in both family based and population-based studies." The editorial says, "In general, we will expect manuscripts reporting genetic associations to include an estimate of the effect size and to contain either a replication in an independent sample or physiologically meaningful data supporting a functional role of the polymorphism in question" (29). Overall, the editors seem to be expressing discomfort in evaluating the quality of epidemiological studies rather than a general objection to population-based studies. A telling point is the final sentence in the editorial: "Our standards will continue to evolve as knowledge improves on. . . appropriate strategies for conducting association studies" (29). If Thomas and Witte want to cite the editorial to support their call for confirmation with family studies, they need to argue more strongly that our work (2) and research on genomic control (13, 23–25) published subsequent to the editorial are not sufficient to lead to changes in the standard.

Cancer Epidemiology Biomarkers and Prevention is publishing this point/counterpoint on population stratification precisely because any differences in practicality, efficiency, and credibility of studies with related and unrelated controls have important ramifications for the design and interpretation of studies on the epidemiology of cancer. If population stratification is truly a serious concern in population-based studies and, therefore, a finding from a family-based study will carry substantially more weight than a finding of similar magnitude and similar apparent precision from a study of unrelated controls, we need to reassess our strategy of using unrelated controls in cohort (22) and case-control (6) studies whose goals are to increase understanding of the genetic component, in combination with environmental factors, of the etiology of complex diseases such as cancer. If, on the other hand, we are correct, and major bias from population stratification is unlikely in well-designed, well-analyzed studies, then funding and publication decisions about cohort and population-based case-control studies should be made on the basis of their adherence to epidemiological design principles and appropriate field-work practices in the context of specific study goals, instead of on the basis of concerns over one source of a small bias.

Without population stratification as a consideration, practical and efficiency considerations dictate the common use of unrelated controls for most, but certainly not all, studies addressing the effects of genetic and environmental factors. We believe that the empirical and theoretical evidence we have presented provides a strong basis for our contention that conclusions from well-designed, well-conducted, and appropriately analyzed and interpreted population-based studies with unrelated controls are robust against bias from population stratification. There is a much more compelling rationale for journals to be wary of low-quality case-control studies than for requiring related controls. There are discrete settings where studies with related controls may help to elucidate the role of genetic factors in the etiology of cancer, but a broad strategy of requiring replication using related controls, out of fear of population stratification or high rates of false positives, is not warranted. Population stratification is not a sufficient reason in itself to reduce reliance on studies using unrelated controls in studies of common polymorphisms and cancer. We reject the principle of the use of related controls as a universal gold standard. They are far from perfect themselves, and provide costly insurance only against the bias from population stratification, which is lilely to be small in most instances.

Acknowledgment

The authors thank Laure el Ghormli for her help with the sample size calculations.

References

1. Thomas, D. C., and Witte, J. S. Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? Cancer Epidemiol. Biomark. Prev., *11:* 2002.

2. Wacholder, S., Rothman, N., and Caporaso, N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J. Natl. Cancer Inst., *92:* 1151–1158, 2000.

3. Wacholder, S., McLaughlin, J. K., Silverman, D. T., and Mandel, J. S. Selection of controls in case-control studies. I. Principles. Am. J. Epidemiol., *135:* 1019–1028, 1992.

4. Thomas, D. B., and Kargas, M. R. Migrant studies. *In:* D. Schottenfeld and J. F. Fraumeni (eds.), Cancer Epidemiology and Prevention, pp. 236–254. New York: Oxford, 1996.

5. Modan, B., Hartge, P., Hirsh-Yechezkel, G., Chetrit, A., Lubin, F., Beller, U., Ben Baruch, G., Fishman, A., Menczer, J., Friedman, E., Piura, B., Ebbers, S. M., Struewing, J. P., Tucker, M. A., and Wacholder, S. Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a *BRCA1* or *BRCA2* mutation. N. Engl. J Med., *345:* 235–240, 2001.

6. Caporaso, N., Rothman, N., and Wacholder, S. Case-control studies of common alleles and environmental factors. J. Natl. Cancer Inst. Monogr., *26:* 25–30, 1999.

7. Rothman, N., Wacholder, S., Caporaso, N., Garcia-Closas, M., Buetow, K., and Fraumeni, J. F. The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. Biochim. Biophys. Acta, *1471:* C1–C10, 2000.

8. Yokoyama, A., Muramatsu, T., Ohmori, T., Kumagai, Y., Higuchi, S., and Ishii, H. Reliability of a flushing questionnaire and the ethanol patch test in screening for inactive aldehyde dehydrogenase-2 and alcohol-related cancer risk. Cancer Epidemiol. Biomark. Prev., *6:* 1105–1107, 1997.

9. Begg, C. B. The search for cancer risk factors: when can we stop looking? Am. J. Public Health, *91:* 360–364, 2001.

10. Risch, N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidemiol. Biomark. Prev., *10:* 733–741, 2001.

11. Risch, N. J. Searching for genetic determinants in the new millennium. Nature (Lond.), *405:* 847–856, 2000.

12. Knowler, W. C., Williams, R. C., Pettitt, D. J., and Steinberg, A. G. $Gm^{3, 5, 13, 14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am. J. Hum. Genet., *43:* 520–526, 1988.

13. Devlin, B., and Roeder, K. Genomic control for association studies. Biometrics, *55:* 997–1004, 1999.

14. Morton, N. E., and Collins, A. Tests and estimates of allelic association in complex inheritance. Proc. Natl. Acad. Sci. USA, *95:* 11389–11393, 1998.

15. Marcus, P. M., Vineis, P., and Rothman, N. *NAT2* slow acetylation and bladder cancer risk: a meta-analysis of 22 case-control studies conducted in the general population. Pharmacogenetics, *10:* 115–122, 2000.

16. Browner, W. S., and Newman, T. B. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. J. Am. Med. Assoc., *257:* 2459–2463, 1987.

17. Garcia-Closas, M., Wacholder, S., Caporaso, N., and Rothman, N. Inference issues in cohort and case-control studies of genetic effects and gene-environment interactions. *In:* M. J. Khoury, J. Little, and W. Burke (eds.), Human Genome Epidemiology: Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease, 2002.

18. Lander, E. S., and Schork, N. J. Genetic dissection of complex traits. Science (Wash. DC), *265:* 2037–2048, 1994.

19. Terwilliger, J. D., and Weiss, K. M. Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr. Opin. Biotechnol., *9:* 578–594, 1998.

20. Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. Nat. Genet., *29:* 306–309, 2001.

21. Wacholder, S., Rothman, N., and Caporaso, N. *Re:* Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. Response. J. Natl. Cancer. Inst., *93:* 157–158, 2001.

22. Langholz, B., Rothman, N., Wacholder, S., and Thomas, D. C. Cohort studies for characterizing measured genes. J. Natl. Cancer Inst. Monogr., *26:* 39–42, 1999.

23. Pritchard, J. K., and Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet., *65:* 220–228, 1999.

24. Reich, D. E., and Goldstein, D. B. Detecting association in a case-control study while correcting for population stratification. Genet. Epidemiol., *20:* 4–16, 2001.

25. Satten, G. A., Flanders, W. D., and Yang, Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am. J. Hum. Genet., *68:* 466–477, 2001.

26. Lin, R. S., and Kelsey, J. E. Use of race and ethnicity in epidemiologic research: concepts, methodological issues, and suggestions for research. Epidemiol. Rev., *22:* 187–202, 2001.

27. Witte, J. S., Gauderman, W. J., and Thomas, D. C. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am. J. Epidemiol., *149:* 693–705, 1999.

28. Teng, J., and Risch, N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. Genome Res., *9:* 234–241, 1999.

29. Freely Associating. Nat. Genet., *22:* 1–2, 1999.